Transforming Facial Weight of Real Images by Editing Latent Space of StyleGAN

V N S Rama Krishna Pinnimty[§], Matt Zhao[§], Palakorn Achananuparp, and Ee-Peng Lim

ABSTRACT

We present an invert-and-edit framework to automatically transform facial weight of an input face image to look thinner or heavier by leveraging semantic facial attributes encoded in the latent space of Generative Adversarial Networks (GANs). Using a pre-trained StyleGAN as the underlying generator, we first employ an optimization-based embedding method to invert the input image into the StyleGAN latent space. Then, we identify the facial-weight attribute direction in the latent space via supervised learning and edit the inverted latent code by moving it positively or negatively along the extracted feature axis. Our framework is empirically shown to produce high-quality and realistic facial-weight transformations without requiring training GANs with a large amount of labeled face images from scratch. Ultimately, our framework can be utilized as part of an intervention to motivate individuals to make healthier food choices by visualizing the future impacts of their behavior on appearance.

I. Introduction

People tend to be less motivated to adopt healthy lifestyles, especially when the consequences of their behaviors are long-term and inconspicuous. For example, young adults who excessively consume diets which are high in calories and added sugar may not develop type-2 diabetes until decades later. Interventions designed to provide information are shown to be ineffective in motivating meaningful behavior change [1], [2]. On the other hand, Appearance-based behavioral interventions, involving emphasizing the impacts of behavior on appearance, have shown to be more effective in motivating behavior changes than traditional information-provision based interventions [1]. Among various forms of appearance-based signals, *facial weight* (also called *facial adiposity*) has proven to be a good predictor of health and health outcomes, including obesity and type-2 diabetes [3]. Thus, an automated tool for simulating changes in facial weight of individuals based on their dietary behavior of future food choices may prove to be a useful feature in an appearance-based dietary intervention.

Existing approaches to transforming the facial weight of face images typically employ computer graphics and image processing techniques, such as 2D face morphing [3] and 3D face reconstruction and face reshaping [4], [5]. Most of these techniques only work well on face images in highly constrained conditions (e.g., frontal pose, neutral expression, and plain background) [3] and often require manual efforts [3], [4] to generate optimal results, making them impractical for societal-scale interventions.

To overcome the problems, we propose a framework based on a recent invert-and-edit approach [6], [7] to incrementally transform the facial weight of a real image by manipulating the image in GAN latent space. By leveraging the latent space of the state-of-the-art StyleGAN [8] which encodes rich semantics of human faces, the proposed framework is able to effectively handle face images with diverse characteristics in both constrained and unconstrained conditions, producing visually compelling transformations without additional manual efforts. Our work contributes to existing research by: (1) exploring the task of progressive/regressive transformation of facial weight by real image editing in latent space; (2) empirically demonstrating the feasibility and practicality of the proposed framework on diverse sets of face images.

II. Related Work

In general, real image manipulations with GANs can be categorized into two major approaches: *image-to-image translation* [9], [10], [11] and *image editing in latent space* [12], [7]. In the former, the goal is to learn the mapping from an input domain to an output domain [9], [10] or all mappings among multiple domains [11]. Whereas, the latter exploits linear interpolations between encoded images in GAN latent space [13], [8] by uncovering and navigating along latent-space directions which correspond to changes in visual attributes. Supervised [6], [7] and self-supervised [12] approaches have been explored to find latent-space directions for image editing.

Our work extends beyond existing face image manipulation work [14], [15], [11], [6], [16], [8], [7] that typically focuses on facial attributes, such as pose, gender, age, expression, bangs, hair color, mouth opening, and eyeglasses. To our knowledge, facial weight progression and regression via latent-space editing has not been explored before. Our framework is built on an *invert-and-edit* approach [17], such as InterFaceGAN [7], which involves obtaining inverted latent codes of input images [18], [19] and uncovering attribute directions in latent space of pre-trained GANs via supervised learning [6], [7]. We further contribute to prior research by empirically evaluating various qualities of the generated images. Lastly, our work and [5] share similar goals, however, their approach is mostly based on 3D face reconstruction and reshaping. Compared to theirs, ours is more extensible, allowing for multi-attribute transformations without retraining.

V.N.S.R.K. Pinnimty, M. Zhao, P. Achananuparp, and E.-P. Lim are with the School of Information Systems, Singapore Management University. Email: {ramap, mattzhao, palakorna, eplim}@smu.edu.sg

[§] Equal contribution

III. Methodology

Our proposed framework leverages the state-of-the-art pre-trained image generation model StyleGAN [8] to create highly realistic facial-weight transformations. Specifically, our goal is to use StyleGAN generator to produce an incremental change in facial weight of an arbitrary input face image from its manipulated latent code.

The framework consists of three main steps. First, the input image is pre-processed to extract and align the face region (see Section III-A). This results in an aligned face image with 1024×1024 resolution. Next, the aligned face image is embedded into the StyleGAN manifold to produce a corresponding latent representation (see Section III-B). We use StyleGAN's extended latent space W^+ for embedding and obtain the inverted latent code as a concatenation of 18 different 512-dimensional *w* vectors. Lastly, we extract a 18×512 dimensional facial-weight attribute vector via supervised learning, algebraically combine the latent code with the extracted attribute vector, and pass the edited latent code to StyleGAN generator to obtain the transformed image (see Section III-C).

A. Pre-processing

In real-world applications, users may supply input images that are not readily suitable for processing by our pipeline. The input image may be of arbitrary resolution, may contain multiple faces, or may not even have any faces. To tackle such issues, we design a robust pre-processing pipeline consisting of the following sequential steps:

Face Detection. We use Max-Margin Object Detection (MMOD) model in the Dlib Python package, which is effective in detecting faces from images even for those with some degree of rotation. When multiple faces are detected, we only keep the primary subject's face (the largest bounding box). The input image will then be cropped around the bounding box of the detected face.

Facial Landmarks Extraction. We use a well-known 68-facial landmark model implemented the Dlib Python package to detect landmarks. Additionally, we use these landmarks to calculate auxiliary parameters like eye-to-eye distance, centroid of the eyes, etc., and compute the angle required for applying face de-rotation.

Face De-rotation. To ensure a high-quality transformation from our pipeline, the input face image is required to be at a near zero degree in-plane rotation. For this, we *warp* and *transform* the input face image to a coordinate space where: faces are centered, eyes lie on a horizontal line, and size of all the faces are approximately identical. Furthermore, we construct a transformation matrix and apply affine transformation to de-rotate the image.

Image Alignment. Lastly, to adjust a given face image into StyleGAN's canonical face position, we apply the same data preparation steps used in Karras et al. [8] for padding, shrinking, or up-scaling. After this step, we get an image with 1024×1024 resolution that is used for latent space embedding.

B. Latent Space Embedding

We adapt the state-of-the-art optimization-based embedding method Image2StyleGAN [18] to invert real images to StyleGAN latent codes. In particular, we further modify the initialization step and the loss functions to improve embedding quality and run-time efficiency. Our embedding algorithm, based on StyleGAN-Encoder [20], is shown in Algorithm 1.

Algorithm 1: Improved Latent Space Embedding

```
Input: Pre-processed image $I \in \mathbb{R}^{1024\times1024\times3}$;
       gradient descent update $F'(.)$; pre-trained
       $ResNet50$ model; learning rate $\eta$
Output: Optimal latent code $w^*$; embedded
        image $G(w)$ optimized via $F'$
$w \leftarrow ResNet50(I)$
$loss_{min} = \infty$
while iteration = $1 \dots E$ do
    $L \leftarrow L_{vgg}(G(w), I) + L_{mse}(G(w), I)$
    $w \leftarrow w - \eta\ F'(\nabla_w.L)$
    if $L < loss_{min}$ then
        $w^* = w$
        $loss_{min} = L$
```

Initialization. First, we start by feeding the aligned face image I to a pre-trained $ResNet50$ [21] model to extract the initial latent code. This latent code, when passed through the StyleGAN generator, gives a corresponding embedded image *G(w)*. We chose $ResNet50$ for its ability in learning better low-level features and its faster rate of convergence over $VGG16$ and $VGG19$. Compared to random initialization or mean-face initialization [18] strategies, this strategy tends to produce inverted latent codes with higher reconstruction quality (see supplementary material) and can quickly converge within the specified number of epochs, achieving a good trade-off between quality and run-time efficiency.

Optimization. Starting from an initial latent code w, we aim to arrive at the closest possible approximation of the input image in the latent space. We perform gradient descent-based optimization using a weighted loss function over a fixed number of iterations. Inspired by [18], we use the same VGG and pixel-wise MSE loss combination with the only difference in the choice of the layers used for calculating the VGG loss. Our proposed loss function is as follows:

$$w^* = arg\ min_w\ \lambda_{vgg} \cdot L_{vgg}(G(w), I)\ +\ \lambda_{mse} \cdot L_{mse}(G(w), I) \quad (1)$$

where w^* is the optimal latent code; λ_{vgg} is the scalar used to assign weight to the VGG perceptual loss; $G(.)$ is the pre-trained StyleGAN generator; w is the latent code to optimize; $I \in \mathbb{R}^{n\times n\times 3}$ is the input image; λ_{mse} is the scalar used to assign weight to the pixel-wise MSE loss.

To get optimal results, we set $\lambda_{vgg} = 1$, $\lambda_{mse} = 1$, $n = 256$ (i.e., $I \in \mathbb{R}^{256\times256\times3}$) when calculating VGG perceptual loss, and $n = 1024$ (i.e., $I \in \mathbb{R}^{1024\times1024\times3}$) when calculating pixel-wise MSE loss.

For VGG loss, we use a single-layer loss involving the $conv3\_2$ layer (layer-9 of VGG16) instead of the multi-layer loss involving $conv1\_1$, $conv1\_2$, $conv3\_2$, and $conv4\_2$ VGG16 layers in the original Image2StyleGAN. Visually, we observed that multi-layer VGG loss did not significantly affect the overall quality of the embedded face images. Formally, our VGG perceptual loss L_{vgg} is defined as follow:

$$L_{vgg}(G(w), I) = \frac{1}{N_9}||F_9(G(w)) - F_9(I)||_2^2 \tag{2}$$

where N_9 is the number of scalars in the output of $conv3\_2$ layer of VGG16; F_9 is the feature output of $conv3\_2$ layer of VGG16.

We use L2-norm for measuring the difference between the pixels. Thus the pixel-wise MSE loss is defined as:

$$L_{mse}(G(w), I) = \frac{1}{N}||G(w) - I||_2^2 \tag{3}$$

where N is the number of scalars in the image (i.e., $N = n \times n \times 3$).

C. Facial-Weight Transformation

The final step in our framework is to manipulate the optimal latent code w^* so that it can be fed into the StyleGAN generator to produce the desired facial-weight transformation. To achieve that, we adopt a general approach similarly employed in [6], [7], which consists of the following steps:

Features Extraction. First, we aim to uncover a hyperplane in the StyleGAN latent space that separates samples into two facial-weight categories, i.e., thin and heavy. This is achieved by training a supervised facial-weight attribute classifier.

We constructed a thin/heavy labeled images dataset by generating 10K synthetic face images along with their latent codes using StyleGAN. After discarding images with noisy artifacts and irregularities, 9.9K images (StyleGAN-9.9K) were kept. Next, each image was manually assigned either a thin (4K) or a heavy (5.9K) class label by one of the co-authors of this paper. Using the manually labeled dataset, we trained a logistic regression classifier to predict a thin/heavy label $\hat{y}$ from a 18×512 dimensional latent code w^*.

$$\hat{y} = f(w^*) = \frac{1}{1 + e^{-(a \cdot w^* + b)}} \tag{4}$$

where a vector parameter a is the desired *facial-weight attribute vector* representing the attribute *direction* in w^*.

As StyleGAN latent space is not perfectly disentangled, manipulating w^* along the facial-weight direction a may inadvertently affect other correlated attributes. To better control the transformation [7], [6], we perform *projection subtraction* to find a projected facial-weight attribute vector $a - proj_x a$ where x is an attribute direction to be disentangled from a. Given n correlated directions; $X = \{x_1, x_2, ..., x_n\}$, we repeat the projection subtraction one direction at a time.

Latent Space Manipulation. Given the projected facial-weight attribute vector, we manipulate the facial-weight attribute of the latent code w^* as follow:

$$w^*_{edit} = w^* + \alpha \cdot a \tag{5}$$

where w^*_{edit} is the edited latent code which when passed through the StyleGAN generator produces transformed images, w^* is the optimal latent code, α is the scalar used to control the degree of transformation towards thinner ($\alpha < 0$) or heavier ($\alpha > 0$) faces, and a is the 18×512 dimensional projected facial-weight attribute vector. We only apply the editing operation to the first 8 layers of w^* as we found them to be the most pertinent layers to facial weight.

IV. Experiments

To measure the performance of our framework, we present experimental evaluations on two sets of face images with varied visual attributes. First, we quantitatively measure: (i) the reconstruction quality of the latent space embedding; and (ii) the visual quality and identity-preserving quality of the transformations. Then, we visually examine examples of transformed images in a qualitative evaluation. Lastly, we assess the realism of the transformations through human evaluation.

A. Experimental Setup

Datasets. We manually selected high-resolution face images of real people from two existing datasets: 100 images from Chicago Face Database (CFD-100) [22] and 100 images from WIDER FACE test set (WIDER-100) [23], as our test datasets.

The original CFD dataset contains images from 597 subjects of Asian, Black, Latino, and White ethnic backgrounds. All CFD images were taken in a constrained condition, i.e., straight frontal pose, neutral facial expression, and plain background. Our CFD-100 samples comprise 30 Asian, 20 Black, 30 Latino, and 20 White subjects with a balanced gender distribution across all groups.

In contrast, the WIDER FACE test images were taken in unconstrained conditions (i.e., "in the wild") with a wide variety of scales, poses, occlusions, expressions, makeups, and illuminations. Additionally, the 100 selected images were manually annotated by one of the co-authors to identify attributes such as gender, age group, ethnicity, facial expression, and angle. The samples consist of subjects with near-uniform splits of genders (50 female and 50 male), age groups (45 young and 55 middle-age or older), ethnicity (55 White and 46 non-White), expressions (34 neutral and 66 non-neutral), and angles (56 frontal and 44 non-frontal). We expect CFD-100 to produce better overall results than WIDER-100.

Implementation Details. For all our experiments, we used StyleGAN trained on 1024×1024 resolution Flickr-Faces-HQ images (StyleGAN-FFHQ) [8].

In the latent space embedding step, we used a pre-trained $ResNet50$ encoder, trained on a dataset of 20k StyleGAN generated face images [20], to obtain the initial latent code.

Next, we used Adam optimizer with the following optimal hyperparameters: learning rate $\eta = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 1e^{-8}$. Moreover, we set the number of iterations $E = 1000$ (Algorithm 1). On average, it took approximately 1.25 minutes to invert one image on a 32GB Nvidia Tesla V100 GPU, compared to 7 minutes when using [18].

Prior to the facial-weight transformation step, we examined potential entanglements between the facial-weight attribute and other facial attributes. Using 200K labeled face images from CelebA dataset [24] with *age*, *gender*, and *mouth-opening expression* attributes, we trained a binary classifier for each attribute. Given the attribute classifiers, we followed similar procedures in Section III-C to uncover the corresponding attribute directions in StyleGAN latent space and measured their correlations with the facial-weight direction using cosine similarity (see supplementary material for details). Next, we performed projection subtraction to disentangle the facial-weight attribute direction from the mouth-opening expression direction (i.e., the most correlated attribute) and used the projected direction for editing.

B. Quantitative Evaluation

Evaluation Metrics. Firstly, to measure the reconstruction quality of the embedded images, we use two standard perceptual metrics: *peak signal-to-noise ratio* (PSNR $\in [0, \infty)$), *structural similarity* (SSIM $\in [0, 1]$), and *perceptual similarity metric* with AlexNet (LPIPS $\in [0, 1]$) [25]. Higher PSNR and SSIM scores suggest better reconstruction quality, whereas higher LPIPS scores indicate lower reconstruction quality. Among these metrics, LPIPS is most consistent with human perception [25]. For each test image, we obtained an aligned image output after the pre-processing step and computed the scores for all 200 aligned-embedded image pairs.

Next, we measure the perceptual quality and the identity preservation aspects of the transformations using *Fréchet Inception Distance* (FID $\in [0, 1]$) and *Openface face recognition* scores (FR $\in [0, 4]$) [26], respectively. In general, lower FID scores indicate higher visual quality. Similarly, lower FR scores suggest that the subject's original identity is more preserved after the transformation. For each dataset, we first generated 5K transformed images from 200 test images using 50 different α values in [-5, 5] range. Then, we computed the FID between a reference set of 200 embedded images and the generated set of 5K images. For FR, we generated four thinner/heavier transformed images using $\alpha = \{-5, -3, 3, 5\}$ for each test image and calculated the scores for 800 aligned-transformed image pairs from both datasets.

Results. We first examine the reconstruction quality by measuring the similarity between the input real images and the embedded images. As shown in Table I, our framework produced better quality embedded images for CFD-100 dataset than WIDER-100 dataset according to all metrics. The results are as expected since CFD-100 data are more visually standardized and less noisy than WIDER-100 data.

TABLE I
QUANTITATIVE EVALUATION RESULTS

	Latent Space Embedding			Transformation	
	PSNR (dB) (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	FR (↓)
CFD-100	32.988	0.764	0.213	15.392	0.218
WIDER-100	31.625	0.747	0.312	33.98	0.392

Next, the mean FID scores in Table I indicate that the transformed CFD-100 images have higher perceptual quality than those of WIDER-100. For identity preservation, we first excluded cases in which aligned images failed to be reconstructed properly by manually checking candidate embedded images with FR $\geq$ 1. As a result, we removed 8 poorly embedded images and 40 corresponding transformed images from WIDER-100. No such failure cases were found in CFD-100 images. After data filtering, the mean FR scores in Table I suggest that the identity of CFD-100 images are more preserved after the transformations than those of WIDER-100. Overall, the results are consistent with our expectation.

C. Qualitative Evaluation

Facial-Weight Transformations. Fig. 1 displays eight selected samples of original 1024×1024 resolution input images (column 1) from CFD-100 (rows 1-4) and WIDER-100 (rows 5-8) datasets and their corresponding embedded (column 4) and transformed images (columns 2-3 and 5-6). As we can see, our framework generates high quality and realistic results. Not only does it produce progressive/regressive changes in specific features (i.e., cheeks, chin, and neck) and facial characteristics (i.e., mouth curvature) during weight gain/loss [3], but it also preserves their identity and ethnicity. These realistic transformations were enabled by the encoded information in StyleGAN latent space without the need for explicit face reshaping functions [5] (see supplementary material for comparisons). Moreover, the latent facial-weight direction is shown to be independent of the natural face shapes. For instance, subjects with a square face shape (rows 6 and 8) were transformed to look heavier without having their faces becoming completely round. Additionally, it is able to generalize for different face angles, e.g., frontal (rows 1-4), 3/4 view (rows 5-7), and upward tilt (row 5). Lastly, a variety of facial expressions are well preserved during the transformation, e.g., neutral (rows 1-4), smiling (rows 5-6), and laughing (row 7). More examples can be found in the supplementary material.

Failure Cases. Fig. 2 shows three examples of common failure cases that highlight the limitations of our framework. The first image depicting a subject with a thinner-transformed ($\alpha = -3$) side-profile face pose suggests the StyleGAN-FFHQ's limit in generalizing beyond frontal and 3/4-view face poses, resulting in facial deformity. One way to handle this out-of-distribution shapes issue is to augment the training data of StyleGAN-FFHQ with more diverse face angles. Next, blob-like artifacts tend to appear in approximately 5% of all generated faces, especially those with a large α value (e.g., $\alpha = -5$ in the second image). This inherent problem has been improved in

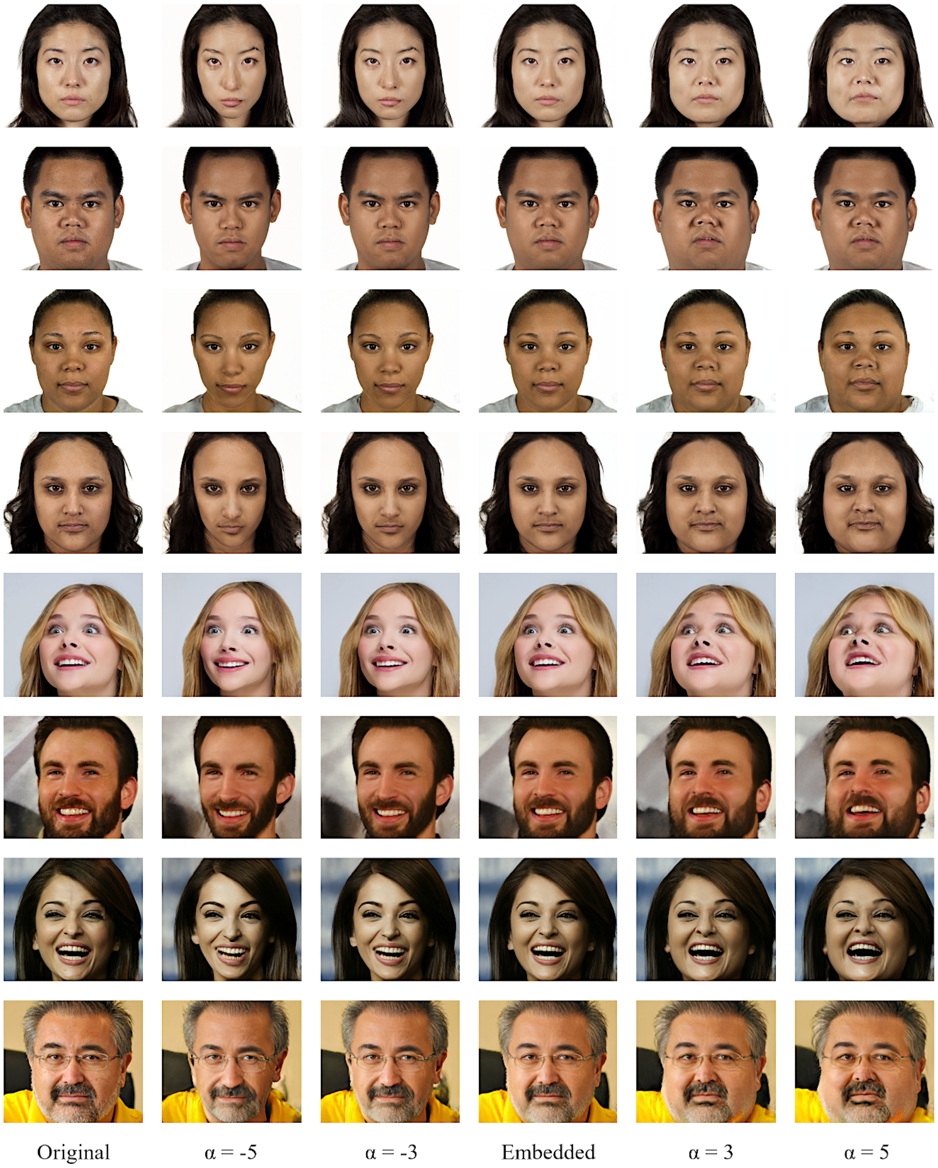

Fig. 1. *Facial-Weight Transformation Results.* Columns 1 and 4 show the original and the embedded images, respectively. Columns 2-3 and 5-6 display the thinner ($\alpha = -5$ and $\alpha = -3$) and heavier ($\alpha = 3$ and $\alpha = 5$) transformations, respectively. Rows 1-4 and 5-8 correspond to samples taken from CFD-100 and WIDER-100 datasets, respectively. Zoom in for better resolution.

StyleGAN2 [27]. The last image shows that, in some small cases, facial weight is still entangled with a mouth-opening expression even though projected attribute vector was already used in editing. Lastly, we observed noticeable artifacts and feature distortions (e.g., elongated or tilted-up noses) in some images, e.g., row 1 in Fig. 1.

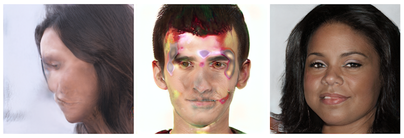

Fig. 2. Examples of common failure cases such as facial deformity, blob artifacts, and entangled features, respectively.

D. Human Evaluation

Setup. We conducted a crowdsourced user study on Amazon Mechanical Turk (AMT) to investigate how humans perceive the realism of our facial-weight transformations. First, we generated thinner ($\alpha = \{-3, -5\}$) and heavier transformations ($\alpha = \{3, 5\}$) for a set of 200 face images (100 StyleGAN-generated and 100 real face images). Given this set of images, we submitted 200 corresponding human intelligence tasks (HITs) to AMT. Each task requires AMT workers to sort a randomly shuffled sequence of five images, i.e., the subject's original image and four of his/her generated images by facial weight from the thinnest to the heaviest (similar to those in Fig. 1). Each task was assigned to three AMT workers, resulting in 600 responses. We further discarded 30 responses due to data input errors made by some workers.

Results. According to 570 crowdsourced responses, our facial-weight transformed images are highly realistic. A vast majority of responses (71.4%) gave the exact ordering of image sequences. In addition, 87.8% of responses correctly identified the thinner transformed images as having lower facial weights than the original subjects. Likewise, 85.2% of responses found the heavier transformed images to be of higher facial weights than the original images. Lastly, small percentages of incorrect responses (13.6% - 15.3%) show the difficulty in distinguishing similar-weight faces, e.g., $\alpha = 3$ vs. $\alpha = 5$.

V. Conclusion

Motivated by appearance-based health intervention, we propose a framework for transforming facial weight of real images by inverting and editing the input images in StyleGAN latent space. Next, we conducted comprehensive experiments to evaluate the performance of our framework using two face images datasets comprising subjects from a diverse demographic backgrounds and visual attributes. The results suggest that not only is our framework capable of producing facial-weight transformed images with high visual quality and realism, it is also effective in preserving the identity and characteristics of subjects after the transformations.

Acknowledgments. This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] R. D. Whitehead, G. Ozakinci, I. D. Stephen, and D. I. Perrett, "Appealing to vanity: could potential appearance improvement motivate fruit and vegetable consumption?" *American Journal of Public Health*, vol. 102, no. 2, pp. 207–11, feb 2012.

[2] P. Achananuparp, E.-P. Lim, and V. Abhishek, "Does journaling encourage healthier choices?: Analyzing healthy eating behaviors of food journalers," in *Proceedings of the 2018 International Conference on Digital Health*, 2018, pp. 35–44.

[3] A. J. Henderson, I. J. Holzleitner, S. N. Talamas, and D. I. Perrett, "Perception of health from facial cues," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 371, no. 1693, 2016.

[4] H. Zhao, X. Jin, X. Huang, M. Chai, and K. Zhou, "Parametric reshaping of portrait images for weight-change," *IEEE Computer Graphics and Applications*, vol. 38, no. 1, pp. 77–90, 2018.

[5] Q. Xiao, X. Tang, Y. Wu, L. Jin, Y.-L. Yang, and X. Jin, "Deep Shapely Portraits," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1800–1808. [Online]. Available: https://doi.org/10.1145/3394171.3413873

[6] S. Guan, "Generating custom photo-realistic faces using ai," Oct 2018. [Online]. Available: https://blog.insightdatascience.com/generating-custom-photo-realistic-faces-using-ai-d170b1b59255

[7] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the Latent Space of GANs for Semantic Face Editing," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-To-Image Translation With Conditional Adversarial Networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jul 2017.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[11] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* IEEE Computer Society, dec 2018, pp. 8789–8797.

[12] A. Jahanian, L. Chai, and P. Isola, "On the "steerability" of generative adversarial networks," in *International Conference on Learning Representations*, 2020.

[13] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015.

[14] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2172–2180.

[15] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. DENOYER, and M. A. Ranzato, "Fader networks:manipulating images by sliding attributes," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5967–5976.

[16] T. Xiao, J. Hong, and J. Ma, "Elegant: Exchanging latent encodings with gan for transferring multiple face attributes," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[17] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation," 2020.

[18] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?" in *Proceedings of the 2019 IEEE International Conference on Computer Vision - ICCV '19*, Oct 2019.

[19] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-Domain GAN Inversion for Real Image Editing," *arXiv e-prints*, p. arXiv:2004.00049, mar 2020.

[20] P. Baylies, "Stylegan-encoder," 2019. [Online]. Available: https://github.com/pbaylies/stylegan-encoder

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of IEEE Computer Vision and Pattern Recognition - CVPR '16*. IEEE, jun 2016, pp. 770–778.

[22] D. S. Ma, J. Correll, and B. Wittenbrink, "The Chicago face database: A free stimulus set of faces and norming data," *Behavior Research Methods*, vol. 47, no. 4, pp. 1122–1135, 2015.

[23] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[25] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[26] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2020.

Appendix

In this section, we present additional analyses and results supplementary to the main paper, including:

- Quantitative and qualitative analyses of the *ResNet50* and mean-face initialization strategies used at the start of the latent space embedding step
- Analysis of potential entanglement between the feature-weight attribute and others
- Additional transformation results
- Comparisons between the results generated by our framework and deep shapely portraits [5]

A. *Initialization Strategies*

We provide additional analysis and examples to compare the reconstruction quality of embedded images using *ResNet50* (used in the main paper) and mean-face (MF) initialization strategies. Quantitatively, the *ResNet50* strategy produces slightly higher LPIPS scores than MF for images from both datasets according to Table II. However, PSNR and SSIM scores of MF are equal or slightly higher than *ResNet50* in both datasets. Given that LPIPS is more consistent with human perceptions than the other metrics [25], We chose *ResNet50* as the initialization strategy in the latent space embedding step.

There are some cases where one strategy was able to generate marginally better results than the other and vice versa. For examples, in Fig. 3, MF is more accurate in reconstructing a ponytail (row 1) and lip shape (row 2) of CFD-100 images than *ResNet50*, whereas *ResNet50* is better at reconstructing some edge cases (rows 3-4) in WIDER-100 than MF.

TABLE II
Reconstruction quality scores for ResNet50 and Mean-Face Initialization Strategies

	PSNR (dB) (↑)	SSIM (↑)	LPIPS (↓)
CFD-100			
ResNet50	32.988	0.764	0.213
Mean-Face	33.023	0.768	0.216
WIDER-100			
ResNet50	31.625	0.747	0.312
Mean-Face	31.580	0.747	0.315

B. *Feature Entanglements*

We provide detailed description on the feature entanglement analysis from which the projected facial-weight vector was derived. In the main paper, we manually annotated 9.9K images (StyleGAN-9.9K) with facial-weight labels in order to extract the facial-weight direction in latent space via supervised learning. To investigate the feature entanglement problem, we measure correlations between the facial-weight direction and other facial attribute features.

To achieve that, we followed similar procedures used in extracting the facial-weight vector. We first trained a binary classifier to predict *age*, *gender*, and *mouth-opening expression* labels (one for each attribute). We selected relevant labeled face images from CelebA dataset [24], resulting in 3 sets of training data; each contains roughly 200K labeled images. For each set, we created a 0.875/0.125 train/test split and trained an attribute classifier by fine-tuning MobileNet (pre-trained on ImageNet). The accuracy scores of age, gender, and mouth-opening classifiers are 0.8863, 0.9353, and 0.8196, respectively. Next, we used the trained classifiers to assign the corresponding labels to StyleGAN-9.9K images. Then, we extracted the three attribute vectors using logistic regression trained on labeled StyleGAN-9.9K images (with 0.7/0.3 train-test split). The accuracy scores of the logistic regression classifiers for facial-weight, age, gender, and mouth-opening expression directions are 0.7993, 0.8034, 0.8312, and 0.7532, respectively.

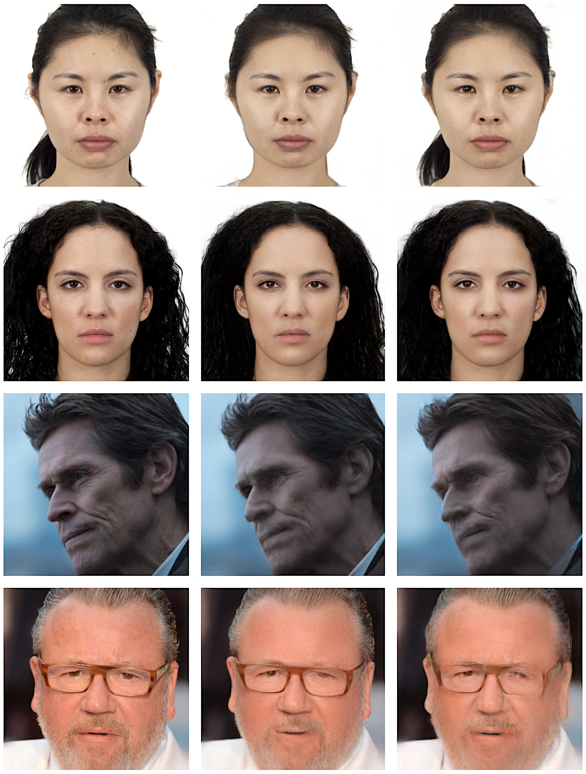

Fig. 3. Comparisons between *ResNet50* and mean-face initialization strategies. Columns 1-3 show the original, *ResNet50*-embedded, and mean-face-embedded images, respectively. Rows 1-2 and 3-4 display images from CFD-100 and WIDER-100 datasets, respectively. Zoom in for better resolution.

TABLE III
Correlations between attribute directions

	Facial weight	Gender	Age	Mouth open
Facial weight	1.000	-0.015	-0.028	0.157
Gender	-	1.000	-0.005	-0.060
Age	-	-	1.000	-0.117
Mouth open	-	-	-	1.000

Finally, we measured cosine similarity between the attribute vectors. As we can see in Table III, mouth-opening expression is more correlated with facial weight than the other attributes. As shown in Fig. 4, the subjects' mouths are more opened when the facial-weight attribute direction was not disentangled with the mouth-opening expression direction (column 3), compared to when projection subtraction was performed (column 2).

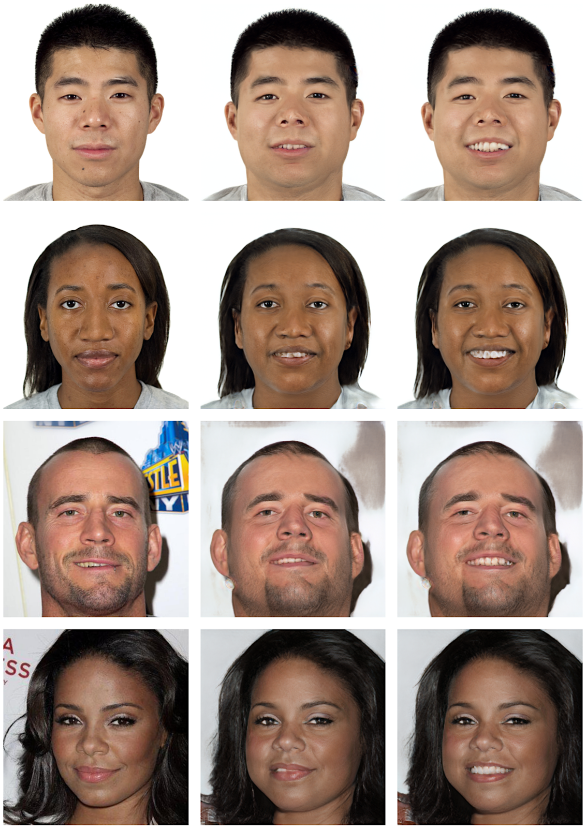

Fig. 4. Feature entanglement with (column 2) and without projected direction (column 3) at $\alpha = 5$ given the input images in column 1

C. Additional Transformation Results

We present additional examples of facial-weight transformation results supplementary to the results in the main paper. Fig. 5-7 and 8-10 display additional transformation results for subjects in CFD-100 and WIDER-100 datasets, respectively. CFD-100 examples aim to show more variations in facial features, face shapes, and body weights within the same ethnicity, whereas WIDER-100 examples illustrate various facial expressions, face angles, and occlusions.

Fig. 11-13 show additional failure cases of face deformities, blob artifacts, and entanglement between facial weight and mouth-opening, respectively. Firstly, we can see in Fig. 11 that face deformities are caused by poor reconstruction quality (rows 1 and 4), occlusions such as hair covering face (row 2) and eyeglasses (row 3), and side-profile poses (rows 5-6). Secondly, Fig. 12 displays occurrences of blob artifacts in different transformation steps, especially when $\alpha < 0$. Lastly, Fig. 13 illustrate variations of mouth-opening expressions that correlate with heavier transformations ($\alpha > 0$). As we can see, navigating along a positive facial weight direction ($\alpha > 0$) sometimes causes the subject's mouth to open slightly.

D. Comparisons with Deep Shapely Portraits

We provide additional results to qualitatively compare our framework with deep shapely portraits (DSP) [5]. DSP is a recent deep-learning based method utilizing sophisticated computer graphics techniques such as 3D face reconstruction, face reshaping, and warping to automatically transform face shapes of portrait images. As we can see in Fig. 14 and 15, the main advantage of DSP (row 2) over our framework (row 3) is in its ability to precisely extract and manipulate face shapes while preserving all other visual elements of the input images (row 1) as it does not recreate the whole image. Even though our framework was not able to accurately reconstruct accessories (e.g., earrings, face tattoo, tassel) and backgrounds without trading computation time for quality, the results demonstrate our framework's effectiveness in preserving the subjects' identity and facial expressions and generating face shapes closely resembling those of DSP, without relying on 3D models and explicit face reshaping functions.

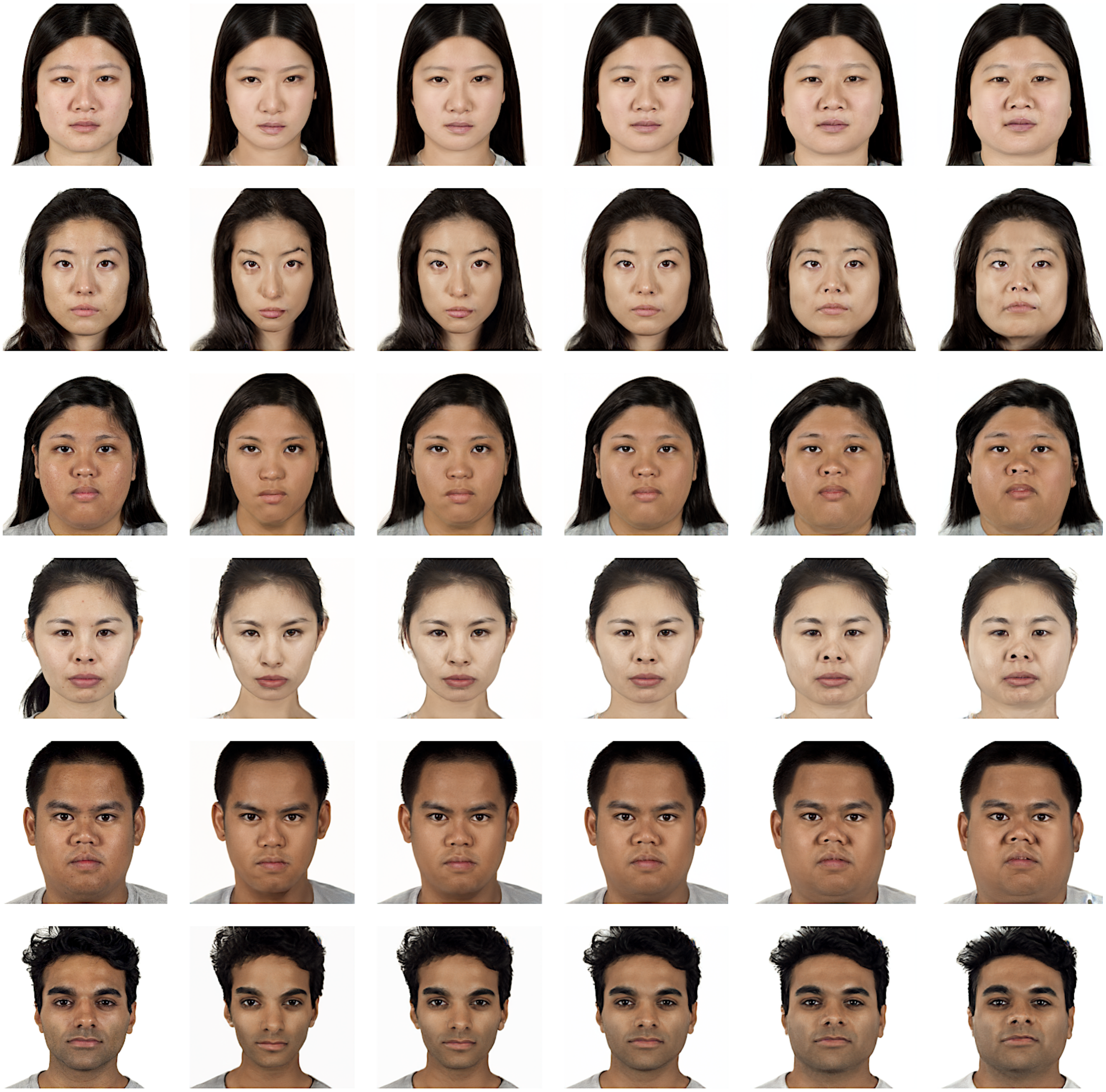

Fig. 5. Additional transformation results for Asian subjects in CFD-100. Columns 1 and 4 show the original and the embedded images, respectively. Columns 2-3 and 5-6 display the thinner ($\alpha = -5$ and $\alpha = -3$) and heavier ($\alpha = 3$ and $\alpha = 5$) transformations, respectively.

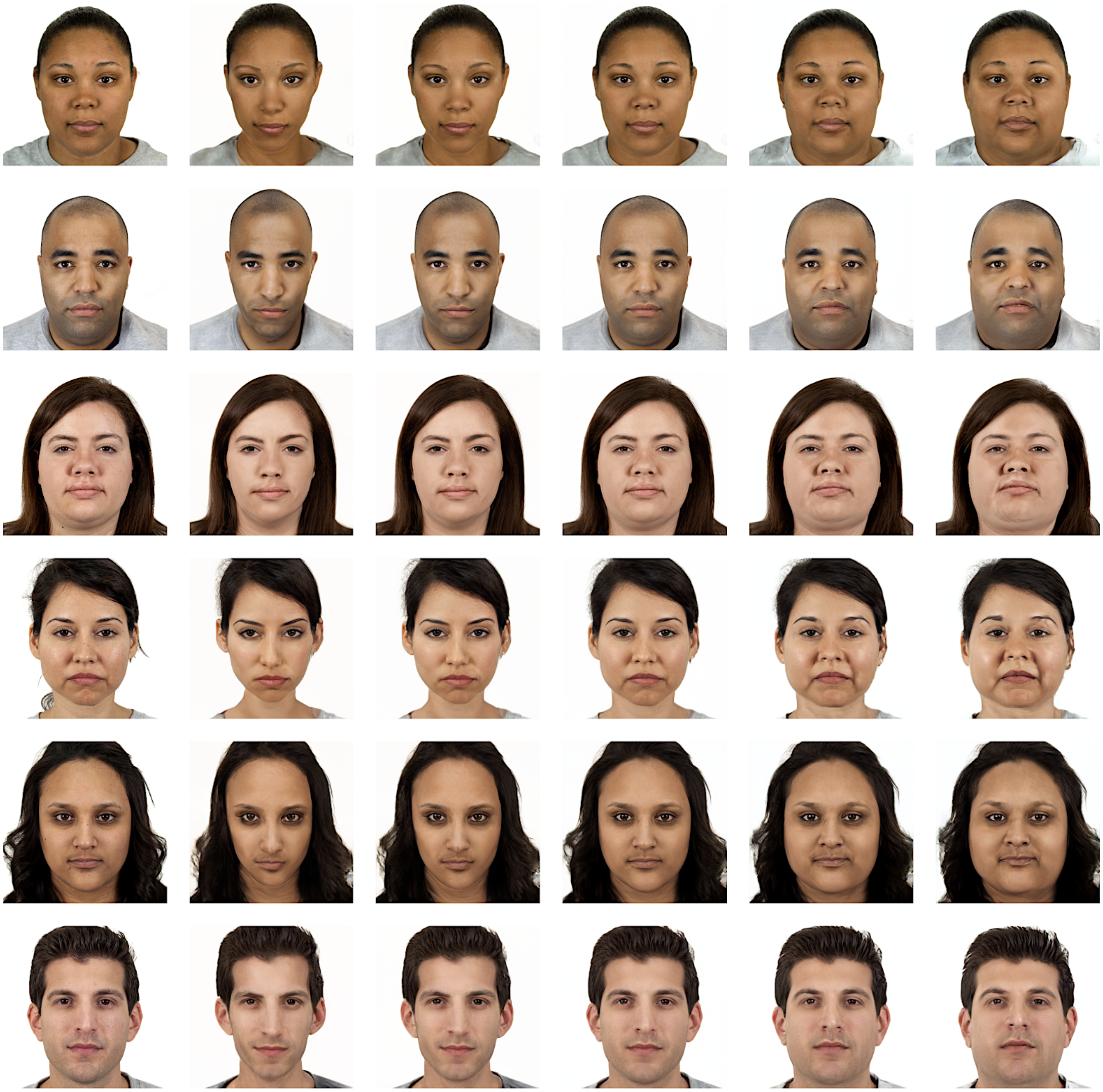

Fig. 6. Additional transformation results for Black and Latino subjects in CFD-100. Columns 1 and 4 show the original and the embedded images, respectively. Columns 2-3 and 5-6 display the thinner ($\alpha = -5$ and $\alpha = -3$) and heavier ($\alpha = 3$ and $\alpha = 5$) transformations, respectively.

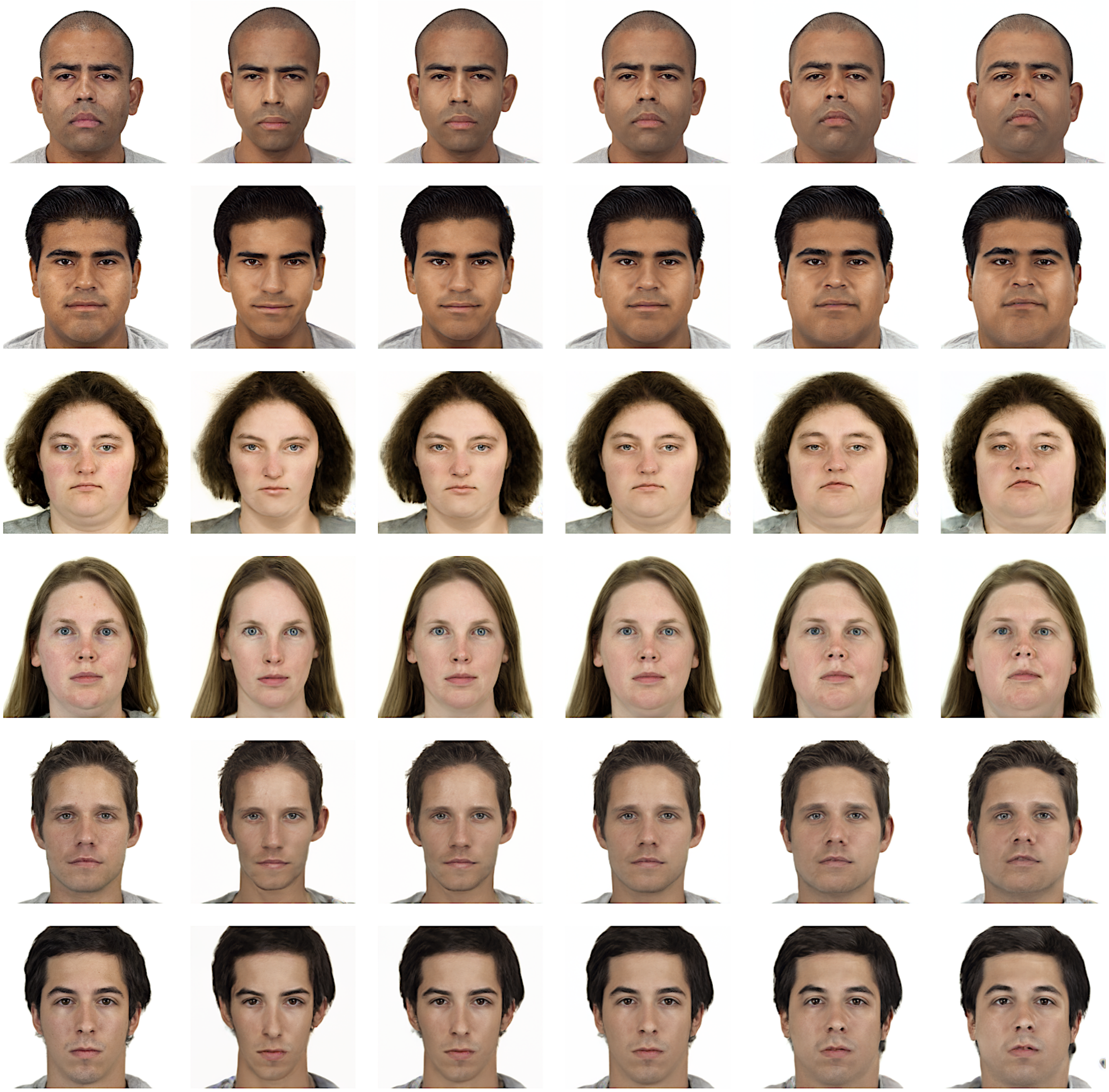

Fig. 7. Additional transformation results for Latino and White subjects in CFD-100. Columns 1 and 4 show the original and the embedded images, respectively. Columns 2-3 and 5-6 display the thinner ($\alpha = -5$ and $\alpha = -3$) and heavier ($\alpha = 3$ and $\alpha = 5$) transformations, respectively.

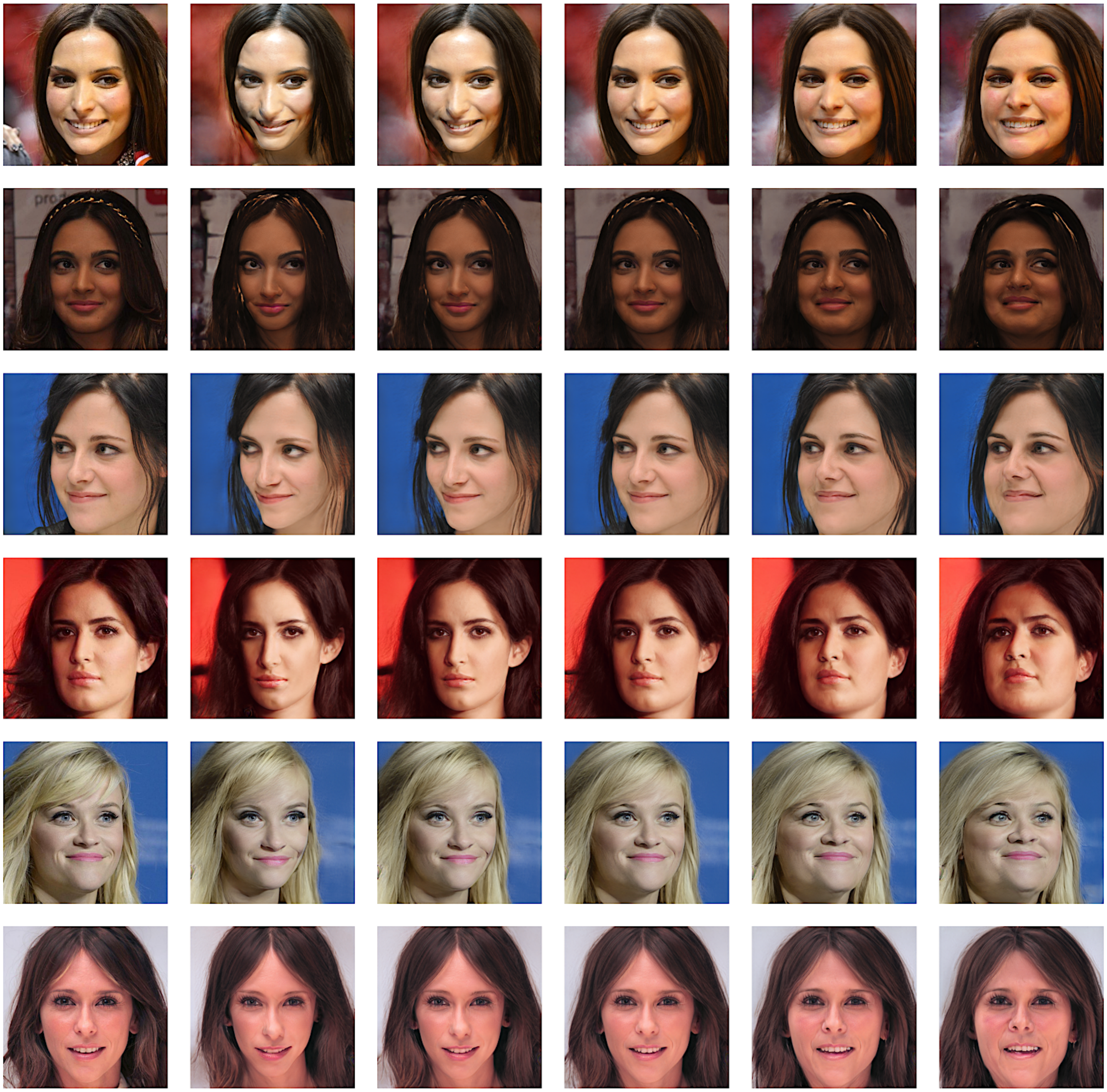

Fig. 8. Additional transformation results for subjects in WIDER-100. Columns 1 and 4 show the original and the embedded images, respectively. Columns 2-3 and 5-6 display the thinner ($\alpha = -5$ and $\alpha = -3$) and heavier ($\alpha = 3$ and $\alpha = 5$) transformations, respectively.

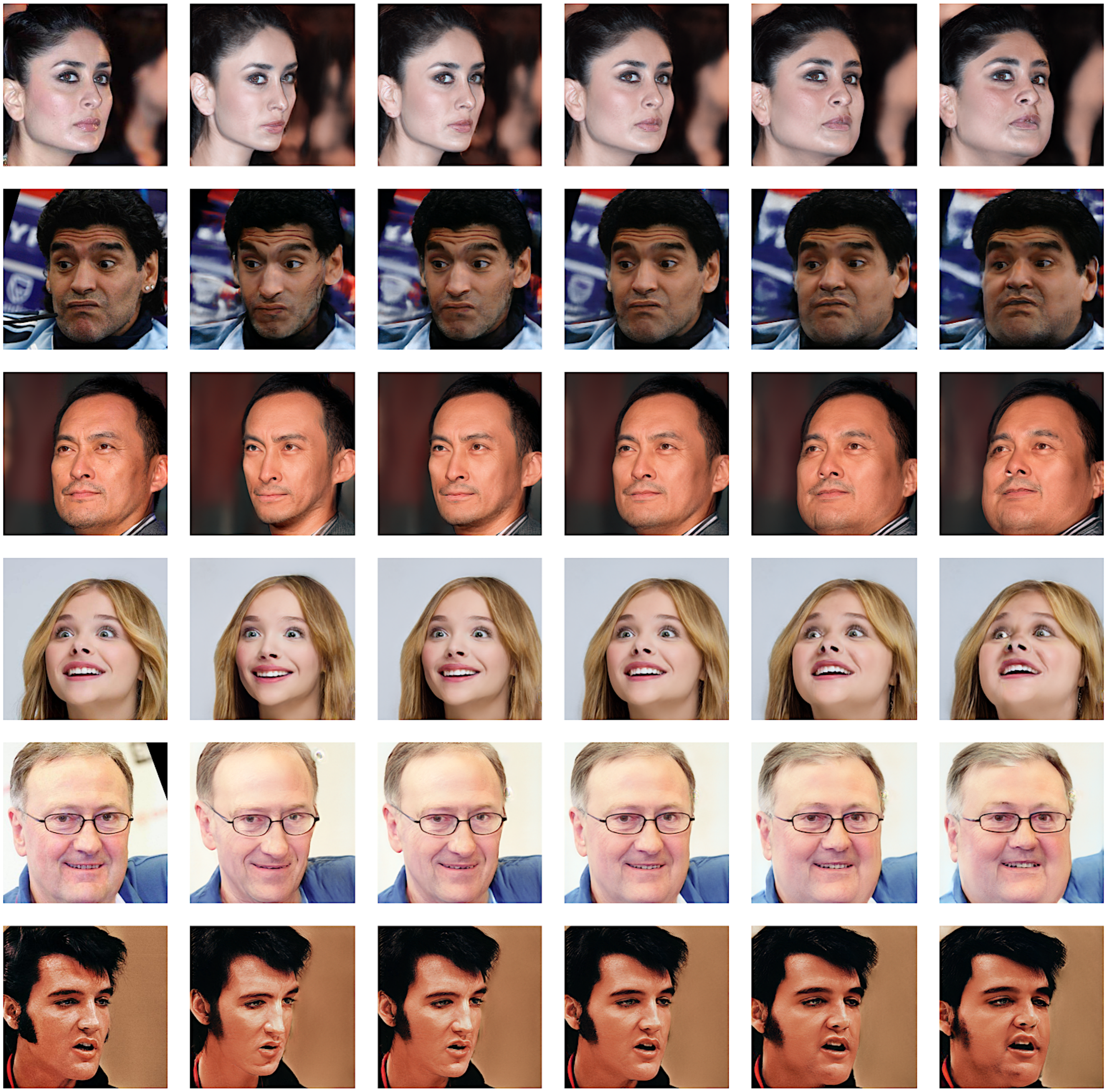

Fig. 9. Additional transformation results for subjects in WIDER-100. Columns 1 and 4 show the original and the embedded images, respectively. Columns 2-3 and 5-6 display the thinner ($\alpha = -5$ and $\alpha = -3$) and heavier ($\alpha = 3$ and $\alpha = 5$) transformations, respectively.

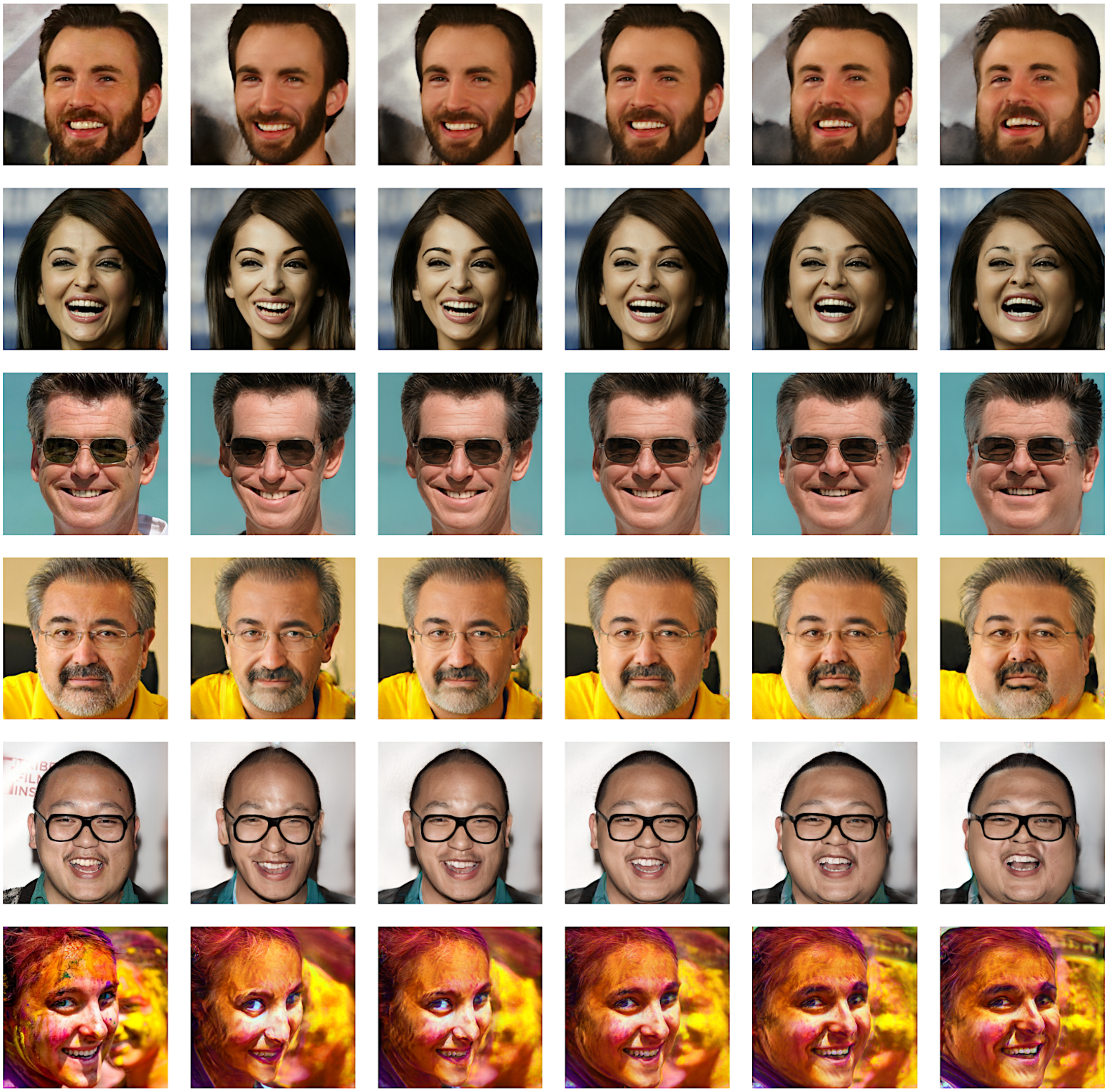

Fig. 10. Additional transformation results for subjects in WIDER-100. Columns 1 and 4 show the original and the embedded images, respectively. Columns 2-3 and 5-6 display the thinner ($\alpha = -5$ and $\alpha = -3$) and heavier ($\alpha = 3$ and $\alpha = 5$) transformations, respectively.

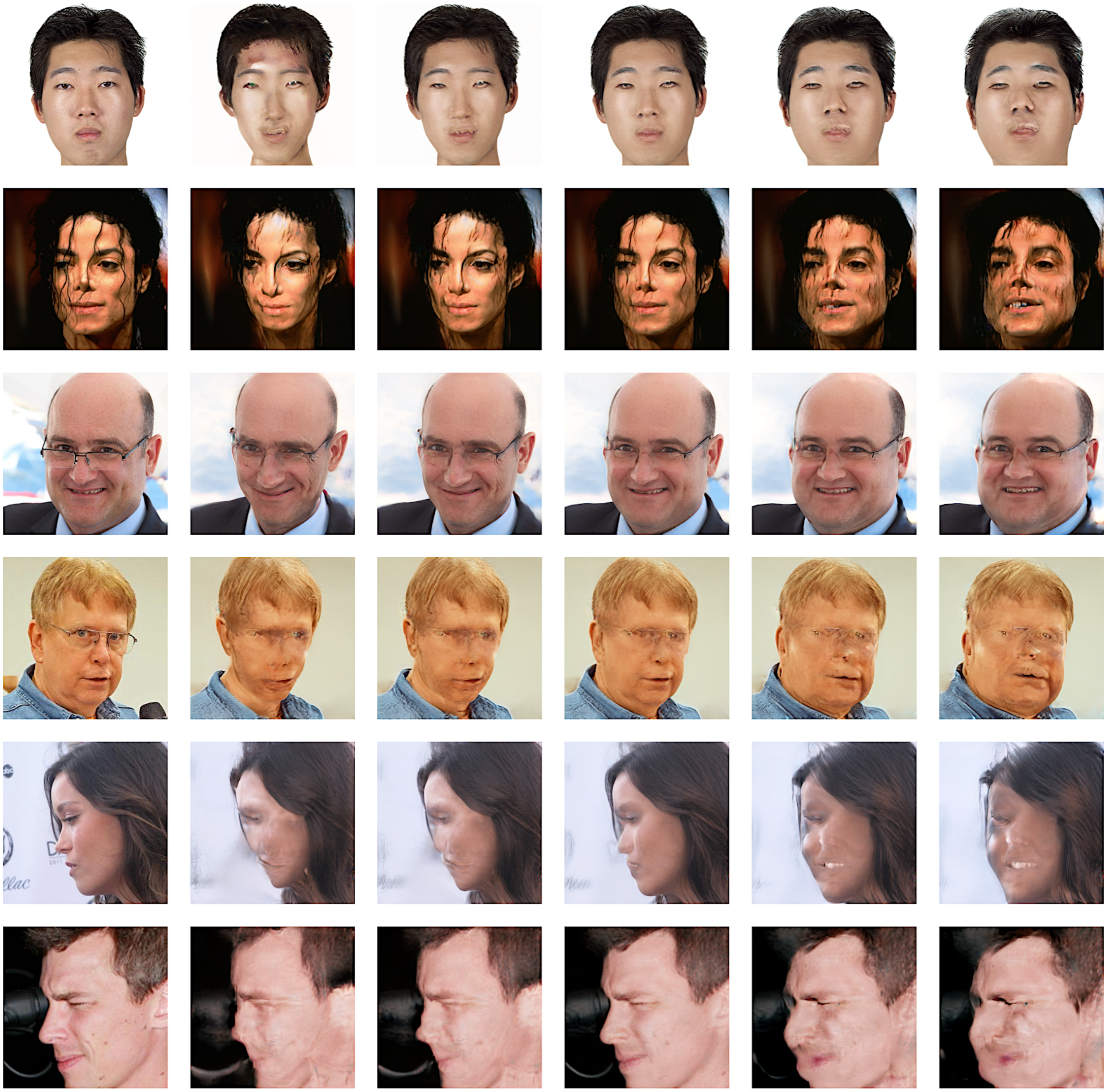

Fig. 11. Additional failure cases of face deformities. Columns 1 and 4 show the original and the embedded images, respectively. Columns 2-3 and 5-6 display the thinner ($\alpha = -5$ and $\alpha = -3$) and heavier ($\alpha = 3$ and $\alpha = 5$) transformations, respectively.

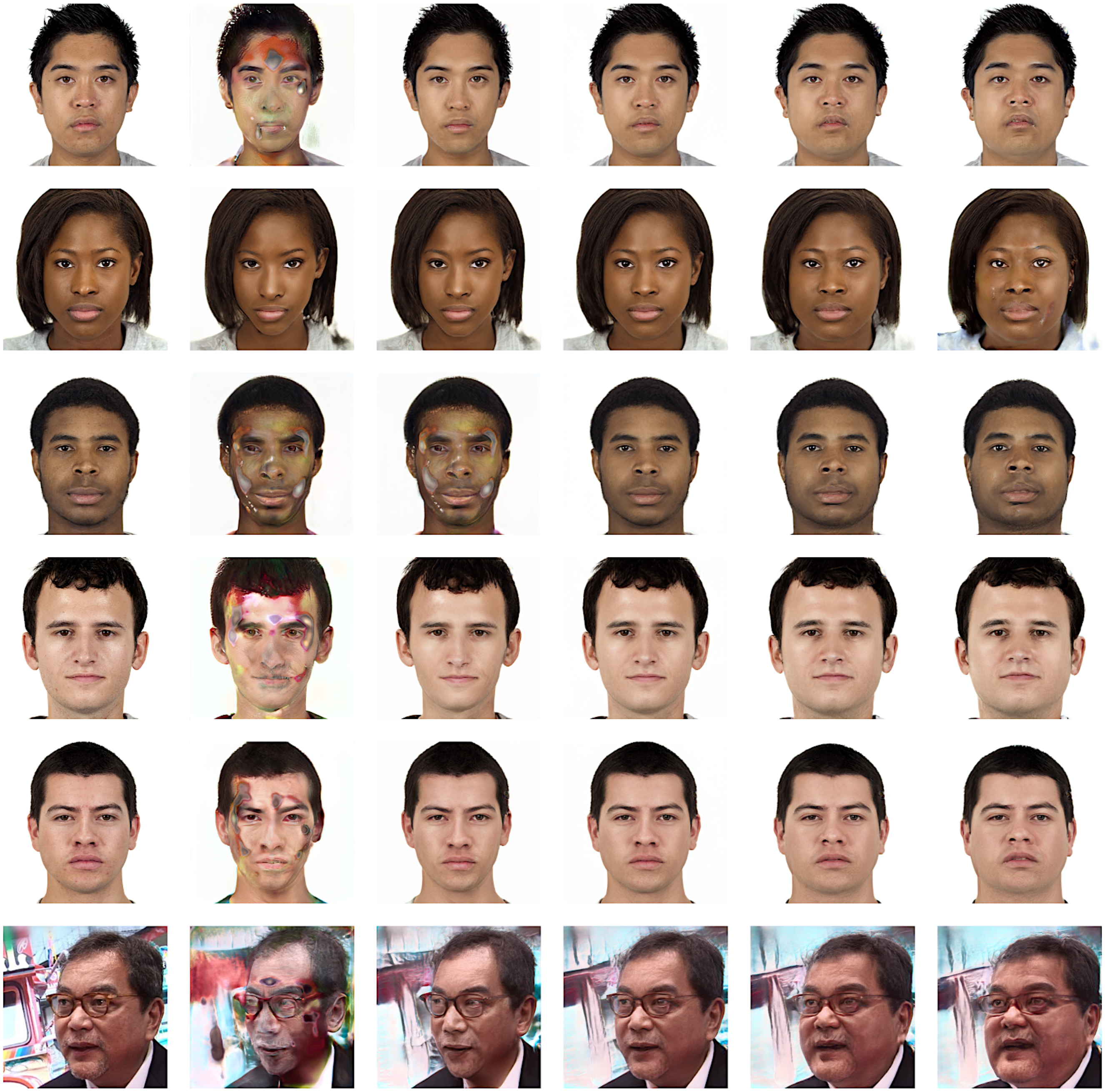

Fig. 12. Additional failure cases of blob artifacts. Columns 1 and 4 show the original and the embedded images, respectively. Columns 2-3 and 5-6 display the thinner ($\alpha = -5$ and $\alpha = -3$) and heavier ($\alpha = 3$ and $\alpha = 5$) transformations, respectively.

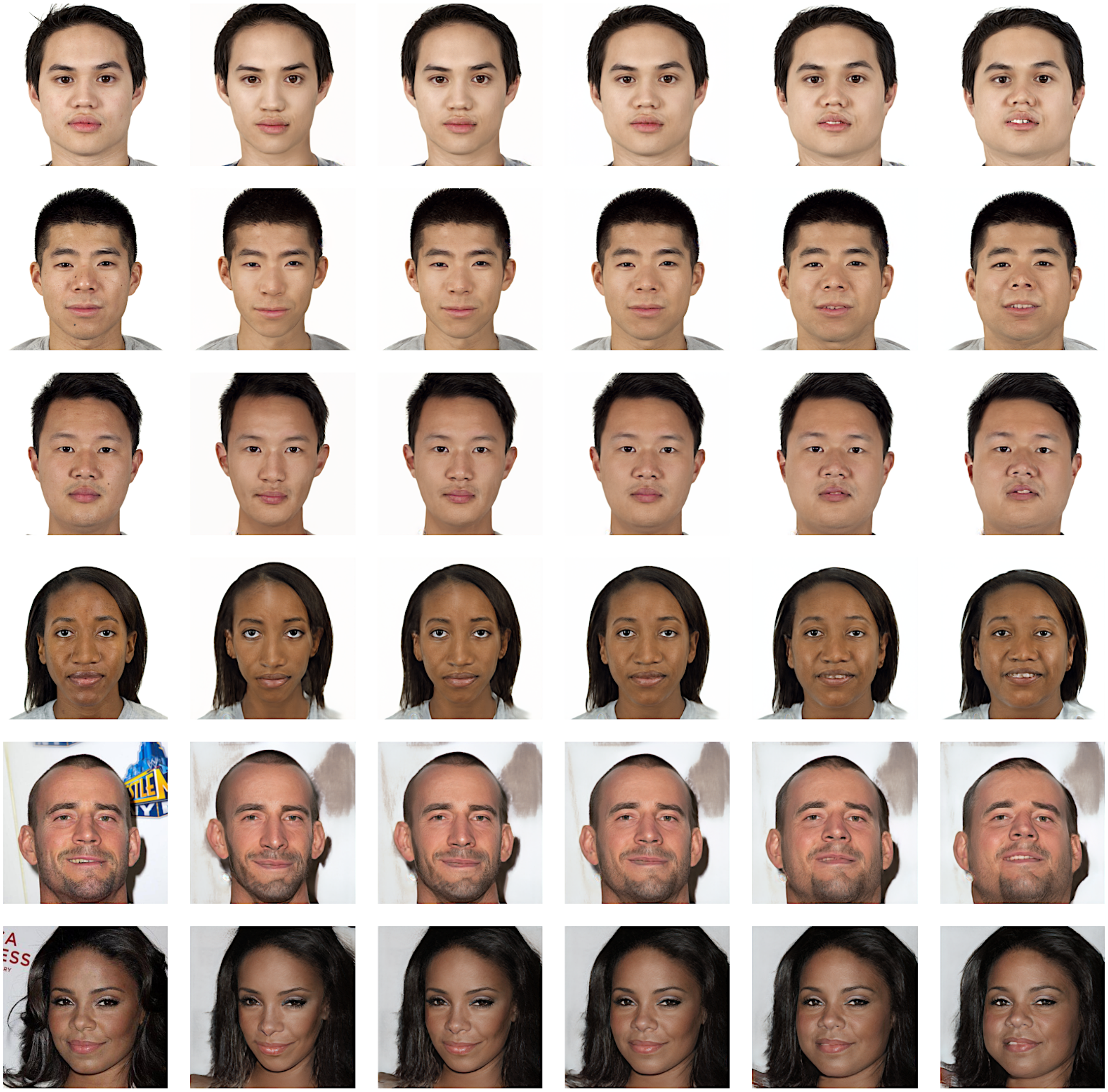

Fig. 13. Additional failure cases of mouth-opening feature entanglement. Columns 1 and 4 show the original and the embedded images, respectively. Columns 2-3 and 5-6 display the thinner ($\alpha = -5$ and $\alpha = -3$) and heavier ($\alpha = 3$ and $\alpha = 5$) transformations, respectively.

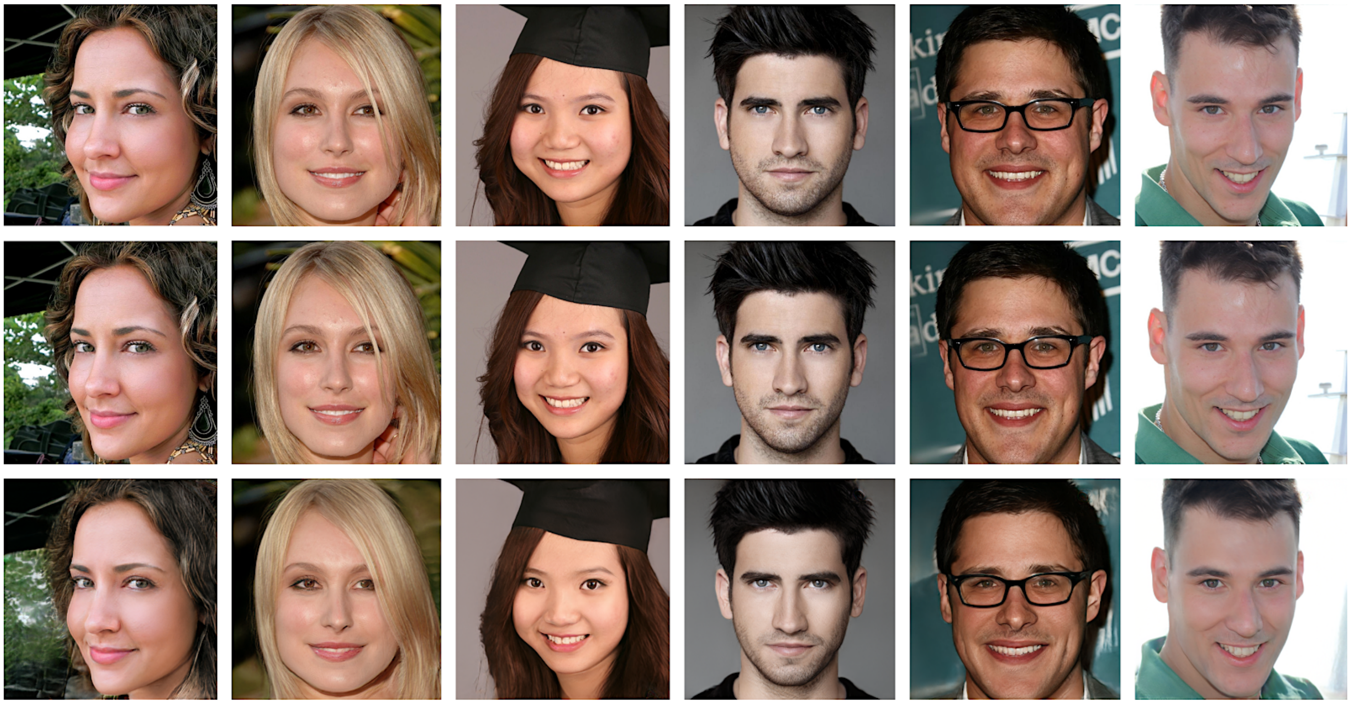

Fig. 14. Comparisons with deep shapely portraits. Columns 1 and 6 show six original images shown in Fig. 1 in Xiao et al.'s study [5]. Rows 1-3 display the original images (row 1), images generated by deep shapely portraits (row 2), and ours (row 3), respectively.

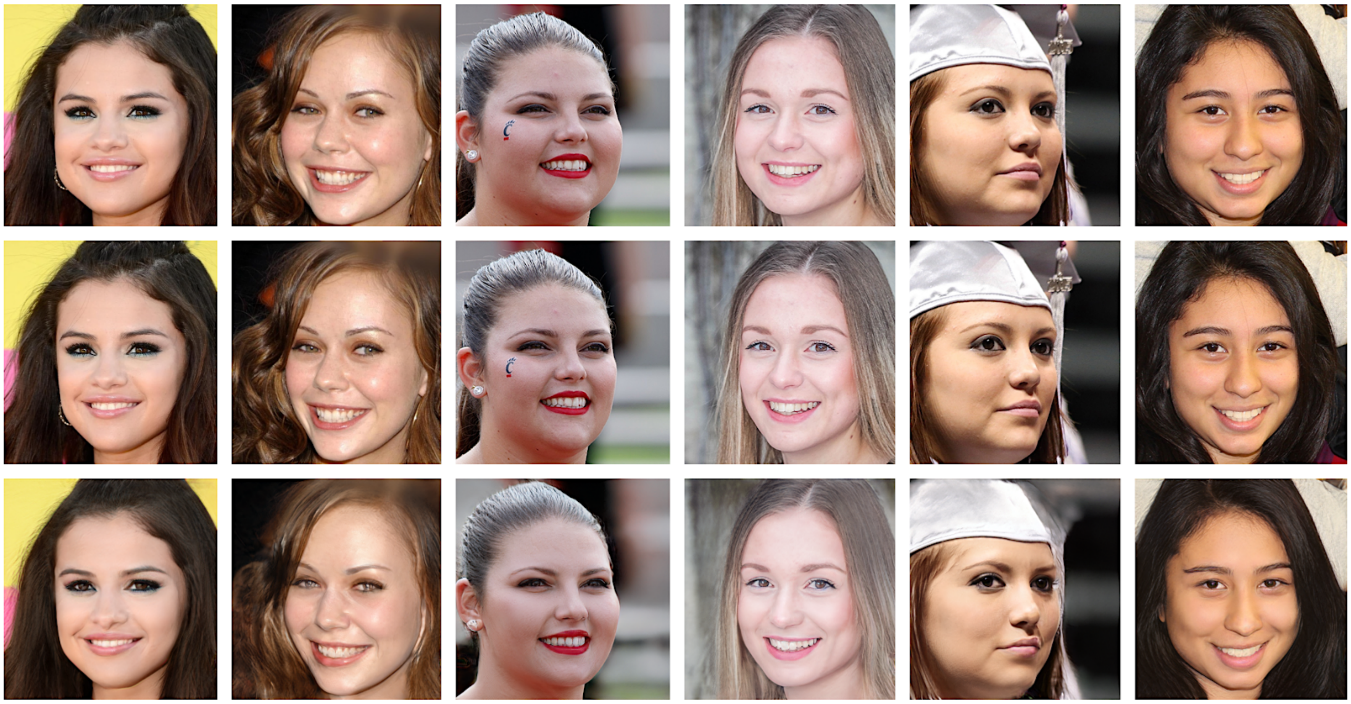

Fig. 15. Comparisons with deep shapely portraits. Columns 1 and 6 show six original images shown in Fig. 6 and 7 in Xiao et al.'s study [5]. Rows 1-3 display the original images (row 1), images generated by deep shapely portraits (row 2), and ours (row 3), respectively.